Detecting COVID-19 Clusters at High Spatiotemporal Resolution, New York City, New York, USA, June–July 2020

Sharon K. Greene, Eric R. Peterson, Dominique Balan, Lucretia Jones, Gretchen M. Culp, Annie D. Fine, Martin Kulldorff

A surveillance system that uses census tract resolution and the SaTScan prospective space-time scan statistic detected clusters of increasing severe acute respiratory syndrome coronavirus 2 test percent positivity in New York City, NY, USA. Clusters included one in which patients attended the same social gathering and another that led to targeted testing and outreach.

Spatiotemporal analysis of high-resolution coronavirus disease (COVID-19) data can help health officials monitor disease spread and target interventions (*1*,*2*). Publicly available data have been used to detect COVID-19 spatiotemporal clusters at county and daily resolution levels across the United States (*3*; R. Amin et al., unpub. data, https://doi.org/10.1101/2020.05.22.20110155) and spatial clusters at ZIP code resolution in New York City (NYC), New York, USA (*4*).

For routine surveillance, the NYC Department of Health and Mental Hygiene (DOHMH) uses the case-only space-time permutation scan statistic (*5*) in SaTScan (https://www.satscan.org) to detect new outbreaks in the context of minimal or stable citywide incidence of reportable diseases (*6*) (e.g., Legionnaires' disease [*7*] and salmonellosis [*8*]). Given wide testing variability, case-only analyses could be poorly suited for COVID-19 monitoring because true differences in disease rates across space and time would be indistinguishable from changing testing rates. We sought to detect in near real-time—regardless whether overall transmission was increasing, decreasing, or steady—newly emerging or re-emerging hotspots (i.e., areas where COVID-19 diagnoses, adjusted for the number of persons tested, were increasing or not decreasing as quickly relative to elsewhere in the city).

Author affiliations: New York City Department of Health and Mental Hygiene, Long Island City, New York, USA (S.K. Greene, E.R. Peterson, D. Balan, L. Jones, G.M. Culp, A.D. Fine); Harvard Medical School, Boston, Massachusetts, USA (M. Kulldorff)

DOI: https://doi.org/10.3201/eid2705.203583

The Study

Clinical and commercial laboratories are required to report all severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) molecular test results (positive, negative, indeterminate) for New York state residents to the New York State Electronic Clinical Laboratory Reporting System (*9*). For NYC residents, this reporting system transmits reports to DOHMH. Laboratory reports include specimen collection date and patient demographics, including residential address, which we geocoded by census tract. Patient symptoms and illness onset date, if any, are obtained through interviews, although not all patients are interviewed.

To detect emerging clusters, the space-time scan statistic uses a cylinder in which the circular base covers a geographic area and the height corresponds to time (*10*). This cylinder is moved, or scanned, over space and time to cover different areas and periods. At each position, the number of cases inside the cylinder is compared with the expected count under the null hypothesis of no clusters by using a likelihood function, and the position with the maximum likelihood is the primary candidate for a cluster. The statistical significance of this cluster is then evaluated, adjusting for the multiple testing inherent in the many cylinder positions evaluated.

To quickly detect emerging hotspots, prospective analyses are conducted daily (*11*). To adjust for the multiple testing stemming from daily analyses, recurrence intervals are used instead of p values (*12*). A recurrence interval of D days means that under the null hypothesis, if we conduct the analysis repeatedly

over D days, then the expected number of clusters of the same or larger magnitude is 1.

The space-time scan statistic can be used with different probability models; we used the Poisson model (*10*), adjusting not for population size (which would fail to account for changing testing rates) but rather for persons tested. Because the goal was to detect newly emerging hotspots rather than areas with consistently high percent positivity, we further adjusted analyses nonparametrically for purely geographic variations that were consistent over time. Fitting a log-linear function, we also adjusted for citywide temporal trends in percent positivity because the goal was to detect local hotspots rather than general citywide trends. For each day and location, the expected count was calculated as the number of persons tested × temporal trend function × a location-specific constant to ensure that, summed over all days in the study period, the location

Table 1. Input file specifications for SARS-CoV-2 test percent positivity cluster detection analyses in New York City, NY, USA, June–July 2020*

Feature	Selection	Notes
Geographic aggregation	Census tract (defined by using US Census 2010 boundaries) of residential address at time of report	With less aggregated data, the more precisely areas with elevated rates can be identified. New York City has 2,165 census tracts located on land. If geocoding is not feasible, then ZIP code could be used but with a loss of spatial precision.
Case file	Unique persons reported with a positive result for a molecular amplification detection (PCR) test for SARS-CoV-2 RNA in a clinical specimen. Retain specimen collection date of first positive test.	Confirmed COVID-19 cases (https://cdn.ymaws.com/www.cste.org/resource/resmgr/2020ps/Interim-20-ID-01_COVID-19.pdf)
Population file	Unique persons reported with a molecular amplification detection (PCR) test for SARS-CoV-2 RNA in a clinical specimen. For persons who ever tested positive, retain specimen collection date of first positive test. Otherwise, retain most recent specimen collection date. For a given census tract and date, if no specimens were collected, then include in file as having 0 population.	Necessary to control for spatial and temporal variability in testing access. A census-based population denominator would not control for variable testing uptake because the number of persons tested is not necessarily proportional to population size.
Omissions from input files	Residents of long-term care facilities, correctional facilities, facilities housing people with developmental disabilities, or homeless shelters; persons whose home address matches selected providers or facilities; persons diagnosed in the 14 d before a more recent case residing in the same building identification number from geocoding; persons with COVID-19 illness onset (where available from patient interview) ≥14 d before specimen collection.	To focus on detecting recent community-based transmission, exclude residents of congregate settings because building-level clusters are detected by using other methods (*13*), persons whose listed home address is not a residence, >1 case/building, and patients whose diagnosis was made long after illness onset.
Date of interest for analysis	Specimen collection date	Defining reportable disease clusters according to when patients became ill is preferred, although a large proportion of COVID-19 infections are asymptomatic. Specimen collection date is the earliest date available for the study population of persons tested.
Study period	21 d for analysis to support prioritization of case investigations; since June 1, 2020, for analysis to support place-based resource allocation	Defining a study period ≥3 times the maximum temporal window helps with statistical power. Extending the study period further may decrease the accuracy of the log-linear temporal trend adjustment but might be of interest for detecting more prolonged clusters. If citywide percent positivity reaches an inflection point (e.g., begins to increase again after a period of decrease), the study period would need to be either temporarily shortened and reset after that inflection point to preserve suitability of a log-linear temporal trend adjustment or a nonparametric temporal trend adjustment could be used. For a longer temporal window, June 1, 2020, was selected as the earliest date when citywide percent positivity trend seemed stable without an inflection point. After 63 d elapsed from June 1, 2020, switched to 63-d rolling study period until next inflection point was reached.
Lag for data accrual	3 d	Given lags between specimen collection and report, exclude very incomplete data at end of study period when estimating the temporal trend. Three days is the minimum lag possible to preserve a timely analysis while allowing for at least some data to be reported, geocoded, and analyzed before open of business.

*The prospective Poisson-based space-time scan statistic was used. COVID-19, coronavirus disease; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Table 2. Spatiotemporal clusters of SARS-CoV-2 test percent positivity prospectively detected and prompting public health action, New York City, NY, USA, June–July 2020*

Maximum temporal window applied, d	Specimen collection date range	Detection date†	Radius, km	Observed cases	Relative risk	Recurrence interval, d	SARS-CoV-2 positivity within cluster, %	Median age (range), y	Notes
7	Jun 17–19	Jun 22	0.6	6	4.0	1	2.2	40 (28–58)	Low recurrence interval; epidemiologic linkage of a gathering identified
21	Jul 5–12	Jul 15	0.6	20	3.4	55	8.9	34 (4–87)	Cluster contributed to selection of area for geographically targeted testing, outreach, and education

*SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.
†To account for data accrual lags, a 3-d delay was imposed between the end of the SaTScan (https://www.satscan.org) study period and the detection date.

has the same number of observed and expected cases. To prioritize quickly emerging clusters to identify epidemiologic linkages, we used a short maximum temporal window of 7 days. To detect sustained clusters to inform place-based resource allocation, starting July 15, we also ran secondary analyses with a maximum temporal window of 21 days.

We developed SAS code (SAS Institute, https://www.sas.com; https://github.com/CityOfNewYork/communicable-disease-surveillance-nycdohmh) to generate daily input and parameter files (Table 1; Appendix Table, https://wwwnc.cdc.gov/EID/article/27/5/20-3583-App1.pdf). The SAS code then invoked SaTScan in batch mode, read analysis results back into SAS for further processing, output files to secured folders (including patient line lists with demographics and map and time-trend visualizations), and sent an investigator notification email.

We launched the system on June 11, 2020, and 2 clusters detected by July 31 prompted public health action (Table 2). First, on June 22, in the context of waning case counts citywide, the only cluster

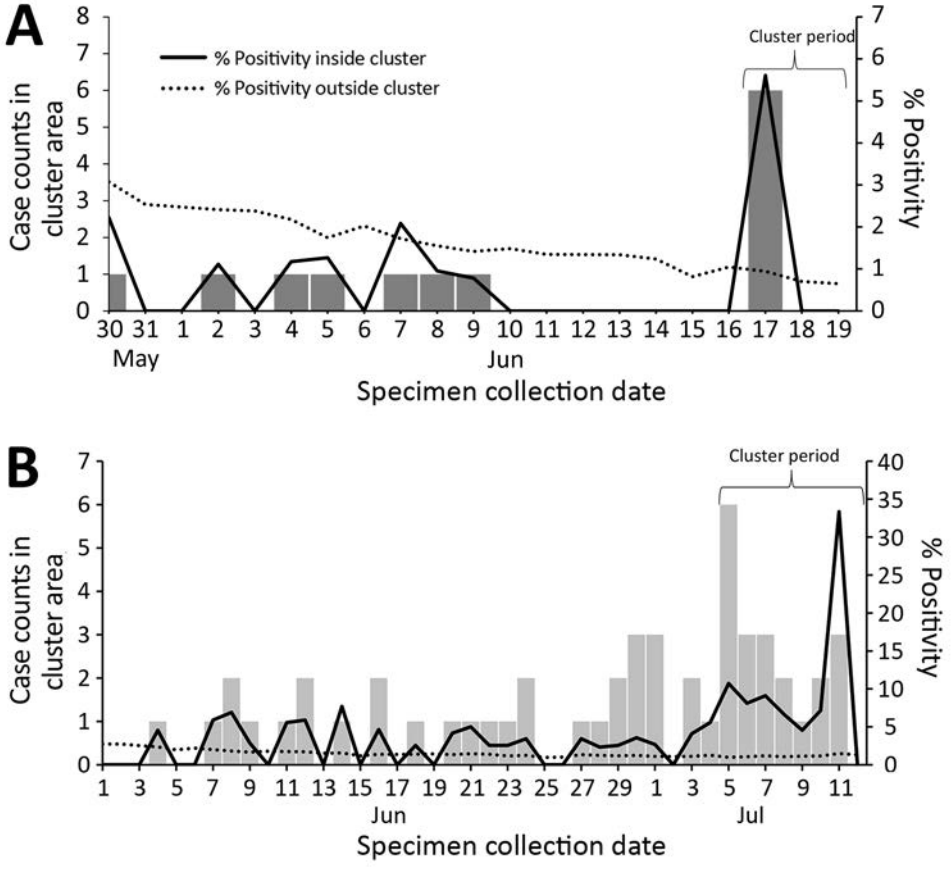

Figure. Cluster case counts and severe acute respiratory syndrome coronavirus 2 test percent positivity inside and outside cluster area for selected clusters detected in New York City, NY, USA, 2020. A) Cluster detected on June 22, 2020, in 5 census tracts in which patients reported common attendance at a social gathering; B) cluster detected on July 15, 2020, in 7 census tracts, contributing to the selection of 1 area for targeted testing and outreach.

detected was of 6 patients residing in a 0.6-km radius, all with specimens collected on June 17 (Figure, panel A). Consequently, DOHMH staff interviewed patients to collect and compare potential common exposures, such as attending the same event or visiting the same location. On June 23, a DOHMH surveillance investigator (D.B.) determined that 2 patients had attended the same gathering, where recommended social distancing practices had not been observed. In response, DOHMH launched an effort to limit further transmission, including testing, contact tracing, community engagement, and health education emphasizing the importance of isolation and quarantine. No other epidemiologic linkages were identified after attempts to investigate ≈65 additional clusters detected through July 2020. Second, detection of a sustained cluster on July 15 (lasting >1 week) with a high percent positivity (Figure, panel B) contributed to geographically targeted testing, outreach, and education, as part of NYC's hyper-local plan to prevent COVID-19 transmission (*14*).

Conclusions

COVID-19 community clusters detected by SaTScan prompted localized public education, testing, and community engagement (*15*). In addition, prioritizing interviews of patients in clusters can identify epidemiologic linkages and opportunities for interrupting further transmission, as is done for other reportable diseases (*6*–*8*). Identification of only 1 linkage in this study could be attributable to changing cluster investigation protocols, low patient response rates, or transmission occurring diffusely in small gatherings. Because all patients are referred for contact tracing, DOHMH discontinued reactively interviewing cluster patients for linkages and instead used clusters to proactively target resources.

The first limitation in this study was timeliness. Analyses were based on specimen collection date; however, given delays in testing availability and care seeking, these dates did not necessarily represent recent infections. Timeliness was further limited by delays from specimen collection to laboratory testing and reporting. Clusters dominated by asymptomatic patients or patients with illness onset >14 days before diagnosis may not require intervention because positive PCR results indicate presence of viral RNA but not necessarily viable virus. The second limitation involved the need to geocode for spatial precision. Of unique NYC residents for whom a specimen was collected for SARS-CoV-2 RNA PCR testing during June–July 2020, residential address was not geocodable for ≈3% of residents, so they were excluded. Third, although recurrence interval thresholds can be used to prioritize responding to clusters (*6*), COVID-19 cluster interpretation can be more complex. Other characteristics for prioritizing COVID-19 clusters, besides statistical significance, include percent positivity, relative risk, case count, epidemic curve trajectory, radius, demographics, and persistence. Prioritization can differ by response activity (e.g., establishing new testing sites, conducting outreach) and how quickly resources can be reallocated. Deciding when and where to initiate interventions in response to COVID-19 clusters cannot be fully automated and requires epidemiologic interpretation.

In summary, our COVID-19 early detection system highlighted areas warranting a rapid response. Targeted, place-based approaches for education and outreach efforts and for localized high transmission warnings could better protect persons at high risk for severe illness and death.

Acknowledgments

We thank all staff members of the DOHMH Incident Command System Surveillance and Epidemiology Section for processing, cleaning, and managing input data; for conducting patient interviews and cluster investigations; and for logistical support. We also thank the NYC Test and Trace Corps for their assistance in managing the cases and contacts included in and identified by cluster investigations.

S.K.G. and E.R.P were supported by the Public Health Emergency Preparedness Cooperative Agreement (grant NU90TP922035-01), funded by the Centers for Disease Control and Prevention (CDC). M.K. was supported by the SaTScan Enhancements Project, managed by the Fund for Public Health in New York City and funded by the CDC Foundation, CDC ELC CARES (grant NU50CK000517-01-09), Alfred P. Sloan Foundation, and Open Society Foundations.

About the Author

Dr. Greene is the director of the Data Analysis Unit at the Bureau of Communicable Disease of the NYC DOHMH, Long Island City, New York. Her research interests include infectious disease epidemiology and applied surveillance methods for outbreak detection.

References

1. De Ridder D, Sandoval J, Vuilleumier N, Stringhini S, Spechbach H, Joost S, et al. Geospatial digital monitoring of COVID-19 cases at high spatiotemporal resolution. Lancet

Digit Health. 2020;2:e393–4. https://doi.org/10.1016/S2589-7500(20)30139-4
2. Furuse Y, Sando E, Tsuchiya N, Miyahara R, Yasuda I, Ko YK, et al. Clusters of coronavirus disease in communities, Japan, January–April 2020. Emerg Infect Dis. 2020;26:2176–9. https://doi.org/10.3201/eid2609.202272
3. Hohl A, Delmelle EM, Desjardins MR, Lan Y. Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. Spat Spatio-Temporal Epidemiol. 2020;34:100354. https://doi.org/10.1016/j.sste.2020.100354
4. Cordes J, Castro MC. Spatial analysis of COVID-19 clusters and contextual factors in New York City. Spat Spatio-Temporal Epidemiol. 2020;34:100355. https://doi.org/10.1016/j.sste.2020.100355
5. Kulldorff M, Heffernan R, Hartman J, Assunção R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS Med. 2005;2:e59. https://doi.org/10.1371/journal.pmed.0020059
6. Greene SK, Peterson ER, Kapell D, Fine AD, Kulldorff M. Daily reportable disease spatiotemporal cluster detection, New York City, New York, USA, 2014–2015. Emerg Infect Dis. 2016;22:1808–12. https://doi.org/10.3201/eid2210.160097
7. Weiss D, Boyd C, Rakeman JL, Greene SK, Fitzhenry R, McProud T, et al.; South Bronx Legionnaires' Disease Investigation Team. A large community outbreak of Legionnaires' disease associated with a cooling tower in New York City, 2015. Public Health Rep. 2017;132:241–50. https://doi.org/10.1177/0033354916689620
8. Latash J, Greene SK, Stavinsky F, Li S, McConnell JA, Novak J, et al. Salmonellosis outbreak detected by automated spatiotemporal analysis—New York City, May–June 2019. MMWR Morb Mortal Wkly Rep. 2020;69:815–9. https://doi.org/10.15585/mmwr.mm6926a2
9. New York State Department of Health. Health advisory: reporting requirements for all laboratory results for SARS-CoV-2, including all molecular, antigen, and serological tests (including "rapid" tests) and ensuring complete reporting of patient demographics [cited 2020 Jun 24]. https://coronavirus.health.ny.gov/system/files/documents/2020/04/doh_covid19_reportingtestresults_rev_043020.pdf
10. Kulldorff M, Athas WF, Feurer EJ, Miller BA, Key CR. Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health. 1998;88:1377–80. https://doi.org/10.2105/AJPH.88.9.1377
11. Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. J R Stat Soc Ser A Stat Soc. 2001;164:61–72. https://doi.org/10.1111/1467-985X.00186
12. Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. Am J Epidemiol. 2004;159:217–24. https://doi.org/10.1093/aje/kwh029
13. Levin-Rector A, Nivin B, Yeung A, Fine AD, Greene SK. Building-level analyses to prospectively detect influenza outbreaks in long-term care facilities: New York City, 2013-2014. Am J Infect Control. 2015;43:839–43. https://doi.org/10.1016/j.ajic.2015.03.037
14. NYC Health + Hospitals. Mayor de Blasio expands hyper-local testing response in Sunset Park, Brooklyn [cited 2020 Nov 27]. https://www.nychealthandhospitals.org/pressrelease/mayor-de-blasio-expands-hyper-local-testing-response-in-sunset-park
15. Stack L, Goldstein J. How a virus surge among Orthodox Jews became a crisis for New York. https://www.nytimes.com/2020/10/08/nyregion/orthodox-jews-queens-brooklyn-closures.html

Address for correspondence: Sharon K. Greene, New York City Department of Health and Mental Hygiene, 42-09 28th St, CN 22A, WS 06-154, Long Island City, NY 11101, USA; email: sgreene4@health.nyc.gov

Article DOI: https://doi.org/10.3201/eid2705.203583

Detecting COVID-19 Clusters at High Spatiotemporal Resolution, New York City, NY, USA, June–July 2020

Appendix

Geocoding

Patient addresses were geocoded daily using version 20A of the NYC Department of City Planning's Geosupport geocoding software, implemented in R through C++ using the Rcpp package (*1*). Addresses that failed to geocode were then cleaned using a string searching algorithm performed against the Department of City Planning's Street Name Dictionary and Property Address Directory. Addresses that failed to geocode after cleaning were then verified using the IBM Infosphere USPS service.

Reference

1. Eddelbuettel D, Francois R. Rcpp: seamless R and C++ integration. J Stat Softw. 2011;40:1–18. https://doi.org/10.18637/jss.v040.i08

Appendix Table. Analysis parameter settings for SARS-CoV-2 test percent positivity cluster detection analyses in New York City, June–July 2020, using the prospective Poisson-based space-time scan statistic.

Parameter	Parameter setting	Notes
Analysis type	Prospective space-time	For timely cluster detection, prospective (rather than retrospective) analyses are used, evaluating only the subset of possible clusters that encompass the last day of the study period. To detect acute, ongoing, localized disease clusters, space-time analyses (rather than purely temporal or purely spatial analyses), are used
Model type	Discrete Poisson	We apply the discrete Poisson-based scan statistic, defining the "population" file as persons tested, to scan for clusters of increased percent positivity. If SARS-CoV-2 percent positivity is high (say, >10%), then the discrete Poisson-based scan statistic is a poor approximation for Bernoulli-type data of persons testing positive and negative. The analysis would produce conservative p-values (i.e., recurrence intervals biased too low). However, for the June–July 2020 period described herein, citywide percent positivity was low, at <4%, so the Poisson model was a very good approximation of the Bernoulli model. Spatial and temporal adjustments for the Bernoulli probability model will be included in a forthcoming SaTScan release and would be preferred in the context of high percent positivity.
Maximum spatial cluster size	50% of the population being tested	The option that imposes the fewest assumptions is to allow the cluster to expand in size to include up to 50% of all persons tested during the study period. Forcing clusters to be smaller than 50%, or restricting in terms of geographic size by setting a maximum circle radius, can be motivated in geographically larger study regions.

Parameter	Parameter setting	Notes
Maximum temporal cluster size	7 d for analysis to support prioritization of case investigations; or 21 d for analysis to support place-based resource allocation	To focus on hotspots emerging during the most recent week; or to focus on areas with more sustained emerging increases.
Minimum temporal cluster size	3 d for analysis to support prioritization of case investigations; or 7 d for analysis to support place-based resource allocation	Clusters of <3-d duration considered less credible for investigation as an emerging hotspot; or clusters of <7 d considered lower priority for resource allocation.
Minimum number of cases	5 cases	Require a minimum number of cases to improve the probability of at least 3 patients within a given cluster being reachable for interview to support identification of a common exposure, or so that resources are not targeted to small numbers of patients.
Temporal trend adjustment	Log-linear with automatically calculated trend	If citywide percent positivity decreasing overall, then wish to detect areas where decreasing slower than citywide average. If citywide percent positivity increasing overall, then wish to detect areas where increasing more than citywide average. A log-linear trend adjustment was suitable for the June–July 2020 period described herein (see percent positivity outside cluster trends in Figure), but when the trend is not log-linear, a different temporal trend adjustment (e.g., log-quadratic, nonparametric) should be used.
Spatial adjustment	Nonparametric, with spatial stratified randomization	The goal during June–July 2020 was to detect areas with relative increases from baseline, even if still lower than average citywide. This method adjusts the expected count separately for each location, removing all purely spatial clusters. The randomization is then stratified by location ID to ensure that each location has the same number of events in the real and random datasets.
Scan for areas with:	High rates	Interested only in increased disease transmission.
Inference	Default p-value method, with maximum number of Monte Carlo replications = 9999	A maximum of 9999 replications increases power compared with 999 replications and is computationally feasible.
Boscoe's limit of clusters by risk level	None	SaTScan can detect clusters with a relative risk near 1, which are not necessarily useful from a public health perspective. It is possible to restrict the analysis to only detect high rate clusters with a minimum relative risk. We did not use this setting during the June–July 2020 period described herein, but in the context of increasing percent positivity, setting a minimum relative risk (e.g., ≥2.0 or 2.5) limits clusters with large spatial extents.
Secondary cluster reporting criteria (output parameter)	No cluster centers in other clusters	COVID-19 may have multiple active clusters at any moment, so secondary clusters should be reviewed. By reviewing clusters with no cluster centers in other clusters (rather than no, or more geographic overlap), secondary clusters with some overlap can be detected. There is no biologically plausible reason to require secondary clusters to have no geographic overlap.